

PATENT APPLICATION  
FOR  
Polymorphisms in the Region of the Human  
Hemochromatosis Gene

Inventors:

DAVID A. RUDDY, a citizen of the United  
States of America, residing at  
855 Greenwich Street  
San Francisco, California 94133;

ROGER K. WOLFF, a citizen of the United  
States of America, residing at  
41 Eugene Street  
Mill Valley, California 94941

Assignee:

Mercator Genetics, Inc.  
4040 Campbell Avenue  
Menlo Park, California 94025;  
a Delaware Corporation.

Entity:

Small Entity

TOWNSEND and TOWNSEND and CREW LLP  
Two Embarcadero Center, 8th Floor  
San Francisco, California 94105  
(415) 326-2400

085495-0009

Polymorphisms in the Region of the Human  
Hemochromatosis Gene

This application is a continuation-in-part of U.S. Patent Application Serial No. 08/724,394, filed October 1, 1996, which is a continuation-in-part of U.S. Patent Application Serial No. 08/630,912, filed April 4, 1996, and U.S. Patent Application Serial No. 08/652,265, filed May 23, 1996, which are herein incorporated by reference in their entirety for all purposes.

BACKGROUND OF THE INVENTION

Hereditary hemochromatosis (HH) is an inherited disorder of iron metabolism wherein the body accumulates excess iron. In symptomatic individuals, this excess iron leads to deleterious effects by being deposited in a variety of organs leading to their failure, and resulting in cirrhosis, diabetes, sterility, and other serious illnesses. The gene which is defective in this disease was disclosed in copending U.S.S.N. 08/652,265.

HH is typically inherited as a recessive trait; in the current state of knowledge, homozygotes carrying two defective copies of the gene are most frequently affected by the disease. In addition, heterozygotes for the HH gene are more susceptible to sporadic porphyria cutanea tarda and potentially other disorders (Roberts et al., Lancet 349:321-323 (1997)). It is estimated that approximately 10-15% of individuals of Northern European descent carry one copy of the HH gene mutation and that there are about one million homozygotes in the United States. HH, thus, represents one of the most common genetic disease mutations in individuals of Northern European descent. Although ultimately HH produces debilitating symptoms, the majority of homozygotes and heterozygotes have not been diagnosed.

The need for such diagnostics is documented, for example, in Barton, J.C. et al. Nature Medicine 2:394-395 (1996); Finch, C.A. West J Med 153:323-325 (1990); McCusick, V. Mendelian Inheritance in Man pp. 1882-1887, 11th ed., (Johns Hopkins University Press, Baltimore (1994)); Report of a Joint World Health Organization/Hemochromatosis Foundation/French Hemochromatosis Association Meeting on the Prevention and Control of Hemochromatosis (1993); Edwards, C.Q. et al. New Engl J Med 328:1616-1620 (1993); Bacon, B.R. New Engl J Med 326:126-127 (1992); Balan, V. et al. Gastroenterology 107:453-459 (1994); Phatak, P.D. et al. Arch Int Med 154:769-776 (1994).

A single mutation in the HH gene, designated 24d1 in pending U.S.S.N. 08/630,912, gave rise to the majority of disease-causing chromosomes present in the population today. This is referred to herein as the "common" or "ancestral" or "common ancestral" mutation. These terms are used interchangeably. It appears that about 80% to 90% of all HH patients carry at least one copy of the common ancestral mutation which is closely linked to specific alleles of certain genetic markers close to this ancestral HH gene defect. These markers are, as a first approximation, in the allelic form in which they were present at the time the ancestral HH mutation occurred. See, for example, Simon, M. et al. Am J Hum Genet 41:89-105 (1987); Jazwinska, E.C. et al. Am J Hum Genet 53:242-257 (1993); Jazwinska, E.C. et al. Am J Hum Genet 56:428-433 (1995); Worwood, M. et al. Brit J Hematol 86:863-866 (1994); Summers, K.M. et al. Am J Hum Genet 45:41-48 (1989).

Several polymorphic markers in the HH region have been described and shown to have alleles that are associated with HH disease. These markers include the published microsatellite markers D6S258, D6S306 (Gyapay, G. et al. Nature Genetics 7:246-339 (1994)), D6S265 (Worwood, M. et al. Brit J Hematol 86:833-846 (1994)), D6S105 (Jazwinska, E.C. et al. Am J Hum Genet 53:242-257 (1993); Jazwinska, E.C. et al. Am J Hum Genet 56:428-433 (1995)), D6S1001 (Stone, C. et al. Hum Molec Genet 3:2043-2046 (1994)), D6S1260 (Raha-Chowdhury

et al. Hum Molec Genet 4:1869-1874 (1995)) as well as additional microsatellite and single-nucleotide-polymorphism markers disclosed in co-pending PCT application WO 96/06583, the disclosure of which is hereby incorporated by reference in its entirety. Additionally, copending U.S.S.N. 08/630,912 disclosed additional markers 24d2 and 24d7.

The symptoms of HH are often similar to those of other conditions, and the severe effects of the disease often do not appear immediately. Accordingly, it would be desirable to provide a method to identify persons who may be destined to become symptomatic in order to intervene in time to prevent excessive tissue damage associated with iron overload. One reason for the lack of early diagnosis is the inadequacy of presently available diagnostic methods to ascertain which individuals are at risk, especially while such individuals are presymptomatic.

Although blood iron parameters can be used as a screening tool, a confirmed diagnosis often employs liver biopsy which is undesirably invasive, costly, and carries a risk of mortality. Thus, there is a clear need for the development of an inexpensive and noninvasive diagnostic test for detection of homozygotes and heterozygotes in order to facilitate diagnosis in symptomatic individuals, provide presymptomatic detection to guide intervention in order to prevent organ damage, and for identification of heterozygote carriers.

#### SUMMARY OF THE INVENTION

One aspect of the invention is an oligonucleotide comprising at least 8 to about 100 consecutive bases from the sequence of Figure 1<sup>(SEQ ID NO: 1)</sup> or Figure 2<sup>(SEQ ID NO: 2)</sup>, or the complement of the sequence, wherein the at least 8 to about 100 consecutive bases includes at least one polymorphic site of Table 1.

Another aspect of the invention is an oligonucleotide pair selected from the sequence of Figure 1<sup>(SEQ ID NO: 1)</sup> or Figure 2<sup>(SEQ ID NO: 2)</sup> or its complement for amplification of a polymorphic site of Table 1.

Another aspect of the invention is an isolated nucleic acid molecule comprising about 100 consecutive bases to about 235 KB, substantially identical to the sequence of Figure 1, <sup>(SEQ ID NO:1)</sup> or Figure 2, <sup>(SEQ ID NO:2)</sup> wherein the DNA molecule comprises at least one polymorphic site of Table 1.

Another aspect of the invention is a method to determine the presence or absence of the common hereditary hemochromatosis (HH) gene mutation in an individual comprising:

providing DNA or RNA from the individual; and assessing the DNA or RNA for the presence or absence of a haplotype of Table 1,

wherein, as a result, the absence of a haplotype of Table 1 indicates the likely absence of the HH gene mutation in the genome of the individual and the presence of the haplotype indicates the likely presence of the HH gene mutation in the genome of the individual.

Another aspect of the invention is a method to determine the presence or absence of the common hereditary hemochromatosis (HH) gene mutation in an individual comprising:

providing DNA or RNA from the individual; and assessing the DNA or RNA for the presence or absence of a genotype defined by a polymorphic allele of Table 1,

wherein, as a result, the absence of a genotype defined by a polymorphic allele of Table 1 indicates the likely absence of the HH gene mutation in the genome of the individual and the presence of the genotype indicates the likely presence of the HH gene mutation in the genome of the individual.

Another aspect of the invention is a culture of lymphoblastoid cells having the designation HC14.

#### BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1, <sup>(SEQ ID NO:1)</sup> depicts the nucleotide sequence of approximately 235 KB in the HH subregion from an unaffected individual.

na

5  
(GPO ID NO. 2)

Figure 2<sub>A</sub> depicts the nucleotide sequence of approximately 235 KB in the HH subregion from an affected individual.

## DETAILED DESCRIPTION

### A. Definitions

Abbreviations for the twenty naturally occurring amino acids follow conventional usage. In the polypeptide notation used herein, the left-hand direction is the amino terminal direction and the right-hand direction is the carboxyl-terminal direction, in accordance with standard usage and convention. Similarly, unless specified otherwise, the left hand end of single-stranded polynucleotide sequences is the 5' end; the left hand direction of double-stranded polynucleotide sequences is referred to as the 5' direction. The direction of 5' to 3' addition of nascent RNA transcripts is referred to as the transcription direction; sequence regions on the DNA strand having the same sequence as the RNA and which are 5' to the 5' end of the RNA transcript are referred to as "upstream sequences"; sequence regions on the DNA strand having the same sequence as the RNA and which are 3' to the 3' end of the RNA transcript are referred to as "downstream sequences".

The term "nucleic acids", as used herein, refers to either DNA or RNA. "Nucleic acid sequence" or "polynucleotide sequence" refers to a single- or double-stranded polymer of deoxyribonucleotide or ribonucleotide bases read from the 5' to the 3' end. It includes both self-replicating plasmids, infectious polymers of DNA or RNA and nonfunctional DNA or RNA. The complement of any nucleic acid sequence of the invention is understood to be included in the definition of that sequence.

"Nucleic acid probes" may be DNA or RNA fragments. DNA fragments can be prepared, for example, by digesting plasmid DNA, or by use of PCR, or synthesized by either the phosphoramidite method described by Beaucage and Carruthers, Tetrahedron Lett. 22:1859-1862 (1981), or by the triester method according to Matteucci, et al., J. Am. Chem. Soc.

103:3185 (1981), both incorporated herein by reference. A double stranded fragment may then be obtained, if desired, by annealing the chemically synthesized single strands together under appropriate conditions or by synthesizing the complementary strand using DNA polymerase with an appropriate primer sequence. Where a specific sequence for a nucleic acid probe is given, it is understood that the complementary strand is also identified and included. The complementary strand will work equally well in situations where the target is a double-stranded nucleic acid.

The phrase "selectively hybridizing to" refers to a nucleic acid probe that hybridizes, duplexes or binds only to a particular target DNA or RNA sequence when the target sequences are present in a preparation of total cellular DNA or RNA. "Complementary" or "target" nucleic acid sequences refer to those nucleic acid sequences which selectively hybridize to a nucleic acid probe. Proper annealing conditions depend, for example, upon a probe's length, base composition, and the number of mismatches and their position on the probe, and must often be determined empirically. For discussions of nucleic acid probe design and annealing conditions, see, for example, Sambrook et al., Molecular Cloning: a Laboratory Manual (2nd ed.), Vols. 1-3, Cold Spring Harbor Laboratory, (1989) or Current Protocols in Molecular Biology, F. Ausubel et al., ed. Greene Publishing and Wiley-Interscience, New York (1987).

The phrase "nucleic acid sequence encoding" refers to a nucleic acid which directs the expression of a specific protein or peptide. The nucleic acid sequences include both the DNA strand sequence that is transcribed into RNA and the RNA sequence that is translated into protein. The nucleic acid sequences include both the full length nucleic acid sequences as well as non-full length sequences derived from the full length protein. It being further understood that the sequence includes the degenerate codons of the native sequence or sequences which may be introduced to provide codon preference in a specific host cell.

The phrase "isolated" or "substantially pure" refers to nucleic acid preparations that lack at least one protein or nucleic acid normally associated with the nucleic acid in a host cell.

5           The phrase "expression cassette", refers to nucleotide sequences which are capable of affecting expression of a structural gene in hosts compatible with such sequences. Such cassettes include at least promoters and optionally, transcription termination signals. Additional factors  
10 necessary or helpful in effecting expression may also be used as described herein.

The term "operably linked" as used herein refers to linkage of a promoter upstream from a DNA sequence such that the promoter mediates transcription of the DNA sequence.

15           The term "vector", refers to viral expression systems, autonomous self-replicating circular DNA (plasmids), and includes both expression and nonexpression plasmids. Where a recombinant microorganism or cell culture is described as hosting an "expression vector," this includes both  
20 extrachromosomal circular DNA and DNA that has been incorporated into the host chromosome(s). Where a vector is being maintained by a host cell, the vector may either be stably replicated by the cells during mitosis as an autonomous structure, or is incorporated within the host's genome.

25           The term "gene" as used herein is intended to refer to a nucleic acid sequence which encodes a polypeptide. This definition includes various sequence polymorphisms, mutations, and/or sequence variants wherein such alterations do not affect the function of the gene product. The term "gene" is  
30 intended to include not only coding sequences but also regulatory regions such as promoters, enhancers, and termination regions. The term further includes all introns and other DNA sequences spliced from the mRNA transcript, along with variants resulting from alternative splice sites.

35           The term "plasmid" refers to an autonomous circular DNA molecule capable of replication in a cell, and includes both the expression and nonexpression types. Where a recombinant microorganism or cell culture is described as



hosting an "expression plasmid", this includes both extrachromosomal circular DNA molecules and DNA that has been incorporated into the host chromosome(s). Where a plasmid is being maintained by a host cell, the plasmid is either being stably replicated by the cells during mitosis as an autonomous structure or is incorporated within the host's genome.

The phrase "recombinant protein" or "recombinantly produced protein" refers to a peptide or protein produced using non-native cells that do not have an endogenous copy of DNA able to express the protein. The cells produce the protein because they have been genetically altered by the introduction of the appropriate nucleic acid sequence. The recombinant protein will not be found in association with proteins and other subcellular components normally associated with the cells producing the protein. The terms "protein" and "polypeptide" are used interchangeably herein.

The following terms are used to describe the sequence relationships between two or more nucleic acids or polynucleotides: "reference sequence", "comparison window", "sequence identity", "percentage of sequence identity", and "substantial identity". A "reference sequence" is a defined sequence used as a basis for a sequence comparison; a reference sequence may be a subset of a larger sequence, for example, as a segment of a full-length cDNA or gene sequence given in a sequence listing, or may comprise a complete cDNA or gene sequence.

Optimal alignment of sequences for aligning a comparison window may, for example, be conducted by the local homology algorithm of Smith and Waterman Adv. Appl. Math. 2:482 (1981), by the homology alignment algorithm of Needleman and Wunsch J. Mol. Biol. 48:443 (1970), by the search for similarity method of Pearson and Lipman Proc. Natl. Acad. Sci. U.S.A. 85:2444 (1988), or by computerized implementations of these algorithms (for example, GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 575 Science Dr., Madison, WI).

The terms "substantial identity" or "substantial sequence identity" as applied to nucleic acid sequences and as

used herein and denote a characteristic of a polynucleotide sequence, wherein the polynucleotide comprises a sequence that has at least 85 percent sequence identity, preferably at least 90 to 95 percent sequence identity, and more preferably at least 99 percent sequence identity as compared to a reference sequence over a comparison window of at least 20 nucleotide positions, frequently over a window of at least 25-50 nucleotides, wherein the percentage of sequence identity is calculated by comparing the reference sequence to the polynucleotide sequence which may include deletions or additions which total 20 percent or less of the reference sequence over the window of comparison. The reference sequence may be a subset of a larger sequence.

As applied to polypeptides, the terms "substantial identity" or "substantial sequence identity" mean that two peptide sequences, when optimally aligned, such as by the programs GAP or BESTFIT using default gap weights, share at least 80 percent sequence identity, preferably at least 90 percent sequence identity, more preferably at least 95 percent sequence identity or more. "Percentage amino acid identity" or "percentage amino acid sequence identity" refers to a comparison of the amino acids of two polypeptides which, when optimally aligned, have approximately the designated percentage of the same amino acids. For example, "95% amino acid identity" refers to a comparison of the amino acids of two polypeptides which when optimally aligned have 95% amino acid identity. Preferably, residue positions which are not identical differ by conservative amino acid substitutions. For example, the substitution of amino acids having similar chemical properties such as charge or polarity are not likely to effect the properties of a protein. Examples include glutamine for asparagine or glutamic acid for aspartic acid.

The phrase "substantially purified" or "isolated" when referring to a peptide or protein, means a chemical composition which is essentially free of other cellular components. It is preferably in a homogeneous state although it can be in either a dry or aqueous solution. Purity and homogeneity are typically determined using analytical

chemistry techniques such as polyacrylamide gel electrophoresis or high performance liquid chromatography. A protein which is the predominant species present in a preparation is substantially purified. Generally, a substantially purified or isolated protein will comprise more than 80% of all macromolecular species present in the preparation. Preferably, the protein is purified to represent greater than 90% of all macromolecular species present. More preferably the protein is purified to greater than 95%, and most preferably the protein is purified to essential homogeneity, wherein other macromolecular species are not detected by conventional techniques.

The phrase "specifically binds to an antibody" or "specifically immunoreactive with", when referring to a protein or peptide, refers to a binding reaction which is determinative of the presence of the protein in the presence of a heterogeneous population of proteins and other biologies. Thus, under designated immunoassay conditions, the specified antibodies bind to a particular protein and do not bind in a significant amount to other proteins present in the sample. Specific binding to an antibody under such conditions may require an antibody that is selected for its specificity for a particular protein. A variety of immunoassay formats may be used to select antibodies specifically immunoreactive with a particular protein. For example, solid-phase ELISA immunoassays are routinely used to select monoclonal antibodies specifically immunoreactive with a protein. See Harlow and Lane (1988) Antibodies, a Laboratory Manual, Cold Spring Harbor Publications, New York, for a description of immunoassay formats and conditions that can be used to determine specific immunoreactivity.

As used herein, "EST" or "Expressed Sequence Tag " refers to a partial DNA or cDNA sequence of about 150 to 500, more preferably about 300, sequential nucleotides of a longer sequence obtained from a genomic or cDNA library prepared from a selected cell, cell type, tissue or tissue type, or organisms which longer sequence corresponds to an mRNA or a gene found in that library. An EST is generally DNA. One or

more libraries made from a single tissue type typically provide at least 3000 different (i.e. unique) EST's and potentially the full complement of all possible EST's representing all possible cDNAs, e.g., 50,000 - 100,000 in an animal such as a human. (See, for example, Adams et al. Science 252:1651-1656 (1991)).

"Stringent" as used herein refers to hybridization and wash conditions of 50% formamide at 42°C. Other stringent hybridization conditions may also be selected. Generally, stringent conditions are selected to be about 5° C lower than the thermal melting point (T<sub>m</sub>) for the specific sequence at a defined ionic strength and pH. The T<sub>m</sub> is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Typically, stringent conditions will be those in which the salt concentration is at least about 0.02 molar at pH 7 and the temperature is at least about 60°C. As other factors may significantly affect the stringency of hybridization, including, among others, base composition and size of the complementary strands, the presence of organic solvents and the extent of base mismatching, the combination of parameters is more important than the absolute measure of any one.

#### B. Polymorphic Markers

The invention provides 397 new polymorphic sites in the region of the HH gene. These polymorphisms are listed in Table 1. As described below, these polymorphisms were identified by comparison of the DNA sequence of an affected individual homozygous for the common ancestral HH mutation with that of an unaffected individual disclosed in copending U.S. 08/724,394.

These polymorphisms provide surrogate markers for use in diagnostic assays to detect the likely presence of the mutations 24d1 and/or 24d2, in preferably 24d1, in homozygotes or heterozygotes. Thus, for example, DNA or RNA from an individual is assessed for the presence or absence of a genotype defined by a polymorphic allele of Table 1, wherein, as a result, the absence of a genotype defined by a

polymorphic allele of Table 1 indicates the likely absence of the HH gene mutation in the genome of the individual and the presence of the genotype indicates the likely presence of the HH gene mutation in the genome of the individual.

These markers may be used singly, in combination with each other, or with other polymorphic markers (such as those disclosed in co-pending PCT application WO 96/06583) in diagnostic assays for the likely presence of the HH gene mutation in an individual. For example, any of the markers defined by the polymorphic sites of Table 1 can be used in diagnostic assays in combination with 24d1 or 24d2, or at least one of polymorphisms HHP-1, HHP-19, or HHP-29, or microsatellite repeat alleles 19D9:205; 18B4:235; 1A2:239; 1E4:271; 24E2:245; 2B8:206; 3321-1:98; 4073-1:182; 4440-1:180; 4440-2:139; 731-1:177; 5091-1:148; 3216-1:221; 4072-2:170; 950-1:142; 950-2:164; 950-3:165; 950-4:128; 950-6:151; 950-8:137; 63-1:151; 63-2:113; 63-3:169; 65-1:206; 65-2:159; 68-1:167; 241-5:108; 241-29:113; 373-8:151; and 373-29:113, D6S258:199, D6S265:122, D6S105:124; D6S306:238; D6S464:206; and D6S1001:180.

Table 2 lists the frequency of about 100 of the alleles defined by the polymorphic sites of the invention in the general population. As is evident from the Table, certain of these alleles are present rarely in the general population. These polymorphisms are thus preferred as surrogate markers in diagnostic assays for the presence of a mutant HH allele ("gene mutation") such as 24d1 or 24d2. Preferably, the frequency of the polymorphic allele used in the diagnostic assay in the general population is less than about 50%, more preferably less than about 25%, and most preferably less than about 5%. Thus, of the genotypes defined by the alleles listed in Table II, polymorphisms occurring at base 35983 and base 61465 of Figure 1 are preferred.

It will be understood by those of skill in the art that because they were identified in an ancestral HH homozygote, the haplotypes defined by the polymorphic sites of Table 1 are predictive of the likely presence of the HH gene mutation 24d1. Thus, for example, the likelihood of any

affected individual having at least two or more of any of the polymorphic alleles defined by Table 1 is greater than that for any unaffected individual. Similarly, the likelihood of any affected individual having at least three or more of any of the polymorphic alleles defined by Table 1 is greater than that for any unaffected individual.

Thus, for example, in a diagnostic assay for the likely presence of the HH gene mutation in the genome of the individual, DNA or RNA from the individual is assessed for the presence or absence of a haplotype of Table 1, wherein, as a result, the absence of a haplotype of Table 1 indicates the likely absence of the HH gene mutation in the genome of the individual and the presence of the haplotype indicates the likely presence of the HH gene mutation in the genome of the individual.

The markers defined by the polymorphic sites of Table 1 are additionally useful as markers for genetic analysis of the inheritance of certain HH alleles and other genes which occur within the chromosomal region corresponding to the sequence of Figure 1, <sup>(SEQ ID NO: 1)</sup> which include, for example, those disclosed in copending U.S.S.N. 08/724,394.

As the entire nucleotide sequence of the region is provided in Figure 1, <sup>(SEQ ID NO: 1)</sup> it will be evident to those of ordinary skill in the art which sequences to use as primers or probes for detecting each polymorphism of interest. Thus, in some embodiments of the invention, the nucleotide sequences of the invention include at least one oligonucleotide pair selected from the sequence of Figure 1, <sup>(SEQ ID NO: 1)</sup> or Figure 2, <sup>(SEQ ID NO: 2)</sup> or its complement for amplification of a polymorphic site of Table 1.

Furthermore, in some embodiments of the invention a preferred hybridization probe is an oligonucleotide comprising at least 8 to about 100 consecutive bases from the sequence of Figure 1, <sup>(SEQ ID NO: 1)</sup> or Figure 2, or the complement of the sequence, wherein the at least 8 to about 100 consecutive bases includes at least one polymorphic site of Table 1. In some embodiments the polymorphic site is at base 35983 or base 61465 of Figure 1, <sup>(SEQ ID NO: 1)</sup>.

It will also be appreciated that the nucleic acid sequences of the invention include isolated nucleic acid

molecules comprising about 100 consecutive bases to about 235  
 KB substantially identical to the sequence of Figure 1<sup>(SEQ ID NO. 1)</sup> or  
 Figure 2<sup>(SEQ ID NO. 2)</sup>, wherein the DNA molecule comprises at least one  
 polymorphic site of Table 1. Such isolated DNA sequences are  
 5 useful as primers, probes, or as the component of a kit in  
 diagnostic assays for detecting the likely presence of the HH  
 gene mutation in an individual.

### C. Nucleic Acid Based Screening

10 Individuals carrying polymorphic alleles of the  
 invention may be detected at either the DNA, the RNA, or the  
 protein level using a variety of techniques that are well  
 known in the art. The genomic DNA used for the diagnosis may  
 be obtained from body cells, such as those present in  
 15 peripheral blood, urine, saliva, bucca, surgical specimen, and  
 autopsy specimens. The DNA may be used directly or may be  
 amplified enzymatically *in vitro* through use of PCR (Saiki et  
 al. Science 239:487-491 (1988)) or other *in vitro*  
 amplification methods such as the ligase chain reaction (LCR)  
 20 (Wu and Wallace Genomics 4:560-569 (1989)), strand  
 displacement amplification (SDA) (Walker et al. Proc. Natl.  
Acad. Sci. U.S.A. 89:392-396 (1992)), self-sustained sequence  
 replication (3SR) (Fahy et al. PCR Methods Appl. 1:25-33  
 (1992)), prior to mutation analysis. The methodology for  
 25 preparing nucleic acids in a form that is suitable for  
 mutation detection is well known in the art.

The detection of polymorphisms in specific DNA  
 sequences, such as in the region of the HH gene, can be  
 accomplished by a variety of methods including, but not  
 30 limited to, restriction-fragment-length-polymorphism detection  
 based on allele-specific restriction-endonuclease cleavage  
 (Kan and Dozy Lancet ii:910-912 (1978)), hybridization with  
 allele-specific oligonucleotide probes (Wallace et al. Nucl  
Acids Res 6:3543-3557 (1978)), including immobilized  
 35 oligonucleotides (Saiki et al. Proc. Natl. Acad. Sci. U.S.A.  
 86:6230-6234 (1989)) or oligonucleotide arrays (Maskos and  
 Southern Nucl Acids Res 21:2269-2270 (1993)), allele-specific  
 PCR (Newton et al. Nucl Acids Res 17:2503-2516 (1989)),

mismatch-repair detection (MRD) (Faham and Cox Genome Res 5:474-482 (1995)), binding of MutS protein (Wagner et al. Nucl Acids Res 23:3944-3948 (1995)), denaturing-gradient gel electrophoresis (DGGE) (Fisher and Lerman et al. Proc. Natl. Acad. Sci. U.S.A. 80:1579-1583 (1983)), single-strand-conformation-polymorphism detection (Orita et al. Genomics 5:874-879 (1983)), RNAase cleavage at mismatched base-pairs (Myers et al. Science 230:1242 (1985)), chemical (Cotton et al. Proc. Natl. Acad. Sci. U.S.A. 85:4397-4401 (1988)) or enzymatic (Youil et al. Proc. Natl. Acad. Sci. U.S.A. 92:87-91 (1995)) cleavage of heteroduplex DNA, methods based on allele specific primer extension (Syvänen et al. Genomics 8:684-692 (1990)), genetic bit analysis (GBA) (Nikiforov et al. Nucl Acids Res 22:4167-4175 (1994)), the oligonucleotide-ligation assay (OLA) (Landegren et al. Science 241:1077 (1988)), the allele-specific ligation chain reaction (LCR) (Barrany Proc. Natl. Acad. Sci. U.S.A. 88:189-193 (1991)), gap-LCR (Abravaya et al. Nucl Acids Res 23:675-682 (1995)), radioactive and/or fluorescent DNA sequencing using standard procedures well known in the art, and peptide nucleic acid (PNA) assays (Orum et al., Nucl. Acids Res. 21:5332-5356 (1993); Thiede et al., Nucl. Acids Res. 24:983-984 (1996)).

In addition to the genotypes defined by the polymorphisms of the invention, as described in co-pending PCT application WO 96/35802 published November 14, 1996, genotypes characterized by the presence of the alleles 19D9:205; 18B4:235; 1A2:239; 1E4:271; 24E2:245; 2B8:206; 3321-1:98 (denoted 3321-1:197 therein); 4073-1:182; 4440-1:180; 4440-2:139; 731-1:177; 5091-1:148; 3216-1:221; 4072-2:170 (denoted 4072-2:148 therein); 950-1:142; 950-2:164; 950-3:165; 950-4:128; 950-6:151; 950-8:137; 63-1:151; 63-2:113; 63-3:169; 65-1:206; 65-2:159; 68-1:167; 241-5:108; 241-29:113; 373-8:151; and 373-29:113, alleles D6S258:199, D6S265:122, D6S105:124, D6S306:238, D6S464:206; and D6S1001:180, and/or alleles associates with the HHP-1, the HHP-19 or HHP-29 single base-pair polymorphisms can also be used to assist in the identification of an individual whose genome contains 24d1 and/or 24d2. For example, the assessing step can be performed



by a process which comprises subjecting the DNA or RNA to amplification using oligonucleotide primers flanking a polymorphism of Table 1, and oligonucleotides flanking 24d1 and/or 24d2, oligonucleotide primers flanking at least one of the base-pair polymorphisms HHP-1, HHP-19, and HHP-29, oligonucleotide primers flanking at least one of the microsatellite repeat alleles, or oligonucleotide primers for any combination of polymorphisms or microsatellite repeat alleles thereof.

Oligonucleotides useful in diagnostic assays are typically at least 8 consecutive nucleotides in length, and may range upwards of 18 nucleotides in length to greater than 100 or more consecutive nucleotides. Such oligonucleotides can be derived from either the genomic DNA of Figure 1 or 2, or cDNA sequences derived therefrom, or may be synthesized. (SEQ ID NO 1) (SEQ ID NO 2)

Additionally, the proteins encoded by such cDNAs are useful in the generation of antibodies for analysis of gene expression and in diagnostic assays, and in the purification of related proteins.

#### D. General Methods

The nucleic acid compositions of this invention, whether RNA, cDNA, genomic DNA, or a hybrid of the various combinations, may be isolated from natural sources, including cloned DNA, or may be synthesized in vitro. The nucleic acids claimed may be present in transformed or transfected whole cells, in a transformed or transfected cell lysate, or in a partially purified or substantially pure form.

Techniques for nucleic acid manipulation of the nucleic acid sequences of the invention such as subcloning nucleic acid sequences encoding polypeptides into expression vectors, labeling probes, DNA hybridization, and the like are described generally in Sambrook et al., Molecular Cloning - a Laboratory Manual (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, (1989), which is incorporated herein by reference. This manual is hereinafter referred to as "Sambrook et al."

There are various methods of isolating the nucleic acid sequences of the invention. For example, DNA is isolated from a genomic or cDNA library using labeled oligonucleotide probes having sequences complementary to the sequences disclosed herein. Such probes can be used directly in hybridization assays. Alternatively probes can be designed for use in amplification techniques such as PCR.

To prepare a cDNA library, mRNA is isolated from tissue such as heart or pancreas, preferably a tissue wherein expression of the gene or gene family is likely to occur. cDNA is prepared from the mRNA and ligated into a recombinant vector. The vector is transfected into a recombinant host for propagation, screening and cloning. Methods for making and screening cDNA libraries are well known. See Gubler, U. and Hoffman, B.J. Gene 25:263-269 (1983) and Sambrook et al.

For a genomic library, for example, the DNA is extracted from tissue and either mechanically sheared or enzymatically digested to yield fragments of about 12-20 KB. The fragments are then separated by gradient centrifugation from undesired sizes and are constructed in bacteriophage lambda vectors. These vectors and phage are packaged *in vitro*, as described in Sambrook, et al. Recombinant phage are analyzed by plaque hybridization as described in Benton and Davis, Science 196:180-182 (1977). Colony hybridization is carried out as generally described in M. Grunstein et al. Proc. Natl. Acad. Sci. USA. 72:3961-3965 (1975).

DNA of interest is identified in either cDNA or genomic libraries by its ability to hybridize with nucleic acid probes, for example on Southern blots, and these DNA regions are isolated by standard methods familiar to those of skill in the art. See Sambrook, et al.

In PCR techniques, oligonucleotide primers complementary to the two 3' borders of the DNA region to be amplified are synthesized. The polymerase chain reaction is then carried out using the two primers. See PCR Protocols: a Guide to Methods and Applications (Innis, M, Gelfand, D., Sninsky, J. and White, T., eds.), Academic Press, San Diego (1990). Primers can be selected to amplify the entire regions

encoding a full-length sequence of interest or to amplify smaller DNA segments as desired.

PCR can be used in a variety of protocols to isolate cDNA's encoding a sequence of interest. In these protocols, appropriate primers and probes for amplifying DNA encoding a sequence of interest are generated from analysis of the DNA sequences listed herein. Once such regions are PCR-amplified, they can be sequenced and oligonucleotide probes can be prepared from sequence obtained.

Oligonucleotides for use as primers or probes are chemically synthesized according to the solid phase phosphoramidite triester method first described by Beaucage, S.L. and Carruthers, M.H., Tetrahedron Lett., 22(20):1859-1862 (1981) using an automated synthesizer, as described in Needham-VanDevanter, D.R., et al., Nucleic Acids Res. 12:6159-6168 (1984). Purification of oligonucleotides is by either native acrylamide gel electrophoresis or by anion-exchange HPLC as described in Pearson, J.D. and Regnier, F.E., J. Chrom., 255:137-149 (1983). The sequence of the synthetic oligonucleotide can be verified using the chemical degradation method of Maxam, A.M. and Gilbert, W., in Grossman, L. and Moldave, D., eds. Academic Press, New York, Methods in Enzymology 65:499-560 (1980).

#### E. Expression

Once DNA encoding a sequence of interest is isolated and cloned, one can express the encoded proteins in a variety of recombinantly engineered cells. It is expected that those of skill in the art are knowledgeable in the numerous expression systems available for expression of DNA encoding a sequence of interest. No attempt to describe in detail the various methods known for the expression of proteins in prokaryotes or eukaryotes is made here.

In brief summary, the expression of natural or synthetic nucleic acids encoding a sequence of interest will typically be achieved by operably linking the DNA or cDNA to a promoter (which is either constitutive or inducible), followed by incorporation into an expression vector. The vectors can

be suitable for replication and integration in either prokaryotes or eukaryotes. Typical expression vectors contain transcription and translation terminators, initiation sequences, and promoters useful for regulation of the expression of polynucleotide sequence of interest. To obtain high level expression of a cloned gene, it is desirable to construct expression plasmids which contain, at the minimum, a strong promoter to direct transcription, a ribosome binding site for translational initiation, and a transcription/translation terminator. The expression vectors may also comprise generic expression cassettes containing at least one independent terminator sequence, sequences permitting replication of the plasmid in both eukaryotes and prokaryotes, *i.e.*, shuttle vectors, and selection markers for both prokaryotic and eukaryotic systems. See Sambrook *et al.* Examples of expression of ATP-sensitive potassium channel proteins in both prokaryotic and eukaryotic systems are described below.

#### 1. Expression in Prokaryotes

A variety of procaryotic expression systems may be used to express the proteins of the invention. Examples include *E. coli*, *Bacillus*, *Streptomyces*, and the like.

It is preferred to construct expression plasmids which contain, at the minimum, a strong promoter to direct transcription, a ribosome binding site for translational initiation, and a transcription/translation terminator. Examples of regulatory regions suitable for this purpose in *E. coli* are the promoter and operator region of the *E. coli* tryptophan biosynthetic pathway as described by Yanofsky, C., J. Bacteriol. 158:1018-1024 (1984) and the leftward promoter of phage lambda ( $P_{\lambda}$ ) as described by Herskowitz, I. and Hagen, D., Ann. Rev. Genet. 14:399-445 (1980). The inclusion of selection markers in DNA vectors transformed in *E. coli* is also useful. Examples of such markers include genes specifying resistance to ampicillin, tetracycline, or chloramphenicol. See Sambrook *et al.* for details concerning selection markers for use in *E. coli*.

To enhance proper folding of the expressed recombinant protein, during purification from *E. coli*, the expressed protein may first be denatured and then renatured. This can be accomplished by solubilizing the bacterially produced proteins in a chaotropic agent such as guanidine HCl and reducing all the cysteine residues with a reducing agent such as beta-mercaptoethanol. The protein is then renatured, either by slow dialysis or by gel filtration. See U.S. Patent No. 4,511,503.

Detection of the expressed antigen is achieved by methods known in the art as radioimmunoassay, or Western blotting techniques or immunoprecipitation. Purification from *E. coli* can be achieved following procedures such as those described in U.S. Patent No. 4,511,503.

## 2. Expression in Eukaryotes

A variety of eukaryotic expression systems such as yeast, insect cell lines, bird, fish, and mammalian cells, are known to those of skill in the art. As explained briefly below, a sequence of interest may be expressed in these eukaryotic systems.

Synthesis of heterologous proteins in yeast is well known. Methods in Yeast Genetics, Sherman, F., et al., Cold Spring Harbor Laboratory, (1982) is a well recognized work describing the various methods available to produce the protein in yeast.

Suitable vectors usually have expression control sequences, such as promoters, including 3-phosphoglycerate kinase or other glycolytic enzymes, and an origin of replication, termination sequences and the like as desired. For instance, suitable vectors are described in the literature (Botstein, et al., Gene 8:17-24 (1979); Broach, et al., Gene 8:121-133 (1979)).

Two procedures are used in transforming yeast cells. In one case, yeast cells are first converted into protoplasts using zymolyase, lyticase or glucanase, followed by addition of DNA and polyethylene glycol (PEG). The PEG-treated protoplasts are then regenerated in a 3% agar medium under

selective conditions. Details of this procedure are given in the papers by J.D. Beggs, Nature (London) 275:104-109 (1978); and Hinnen, a., et al., Proc. Natl. Acad. Sci. U.S.A. 75:1929-1933 (1978). The second procedure does not involve  
5 removal of the cell wall. Instead the cells are treated with lithium chloride or acetate and PEG and put on selective plates (Ito, H., et al., J. Bact. 153:163-168 (1983)).

The proteins of the invention, once expressed, can be isolated from yeast by lysing the cells and applying  
10 standard protein isolation techniques to the lysates. The monitoring of the purification process can be accomplished by using Western blot techniques or radioimmunoassay of other standard immunoassay techniques.

The sequences encoding the proteins of the invention  
15 can also be ligated to various expression vectors for use in transforming cell cultures of, for instance, mammalian, insect, bird or fish origin. Illustrative of cell cultures useful for the production of the polypeptides are mammalian cells. Mammalian cell systems often will be in the form of  
20 monolayers of cells although mammalian cell suspensions may also be used. A number of suitable host cell lines capable of expressing intact proteins have been developed in the art, and include the HEK293, BHK21, and CHO cell lines, and various human cells such as COS cell lines, HeLa cells, myeloma cell  
25 lines, Jurkat cells, etc. Expression vectors for these cells can include expression control sequences, such as an origin of replication, a promoter (e.g., the CMV promoter, a HSV *tk* promoter or *pgk* (phosphoglycerate kinase) promoter), an enhancer (Queen et al. Immunol. Rev. 89:49 (1986)), and  
30 necessary processing information sites, such as ribosome binding sites, RNA splice sites, polyadenylation sites (e.g., an SV40 large T Ag poly a addition site), and transcriptional terminator sequences. Other animal cells useful for  
production of ATP-sensitive potassium channel proteins are  
35 available, for instance, from the American Type Culture Collection Catalogue of Cell Lines and Hybridomas (7th edition, (1992)).

Appropriate vectors for expressing the proteins of the invention in insect cells are usually derived from the SF9 baculovirus. Suitable insect cell lines include mosquito larvae, silkworm, armyworm, moth and *Drosophila* cell lines such as a Schneider cell line (See Schneider J. Embryol. Exp. Morphol. 27:353-365 (1987)).

As indicated above, the vector, e.g., a plasmid, which is used to transform the host cell, preferably contains DNA sequences to initiate transcription and sequences to control the translation of the protein. These sequences are referred to as expression control sequences.

As with yeast, when higher animal host cells are employed, polyadenylation or transcription terminator sequences from known mammalian genes need to be incorporated into the vector. An example of a terminator sequence is the polyadenylation sequence from the bovine growth hormone gene. Sequences for accurate splicing of the transcript may also be included. An example of a splicing sequence is the VP1 intron from SV40 (Sprague, J. et al., J. Virol. 45: 773-781 (1983)).

Additionally, gene sequences to control replication in the host cell may be incorporated into the vector such as those found in bovine papilloma virus type-vectors. Saveria-Campo, M., 1985, "Bovine Papilloma virus DNA a Eukaryotic Cloning Vector" in DNA Cloning Vol. II a Practical Approach Ed. D.M. Glover, IRL Press, Arlington, Virginia pp. 213-238.

The host cells are competent or rendered competent for transformation by various means. There are several well-known methods of introducing DNA into animal cells. These include: calcium phosphate precipitation, fusion of the recipient cells with bacterial protoplasts containing the DNA, treatment of the recipient cells with liposomes containing the DNA, DEAE dextran, electroporation and micro-injection of the DNA directly into the cells.

The transformed cells are cultured by means well known in the art. Biochemical Methods in Cell Culture and Virology, Kuchler, R.J., Dowden, Hutchinson and Ross, Inc., (1977). The expressed polypeptides are isolated from cells

grown as suspensions or as monolayers. The latter are recovered by well known mechanical, chemical or enzymatic means.

5 D. Purification

The proteins produced by recombinant DNA technology may be purified by standard techniques well known to those of skill in the art. Recombinantly produced proteins can be directly expressed or expressed as a fusion protein. The  
10 protein is then purified by a combination of cell lysis (e.g., sonication) and affinity chromatography. For fusion products, subsequent digestion of the fusion protein with an appropriate proteolytic enzyme releases the desired polypeptide.

The polypeptides of this invention may be purified  
15 to substantial purity by standard techniques well known in the art, including selective precipitation with such substances as ammonium sulfate, column chromatography, immunopurification methods, and others. See, for instance, R. Scopes, Protein Purification: Principles and Practice, Springer-Verlag: New  
20 York (1982), incorporated herein by reference. For example, antibodies may be raised to the proteins of the invention as described herein. Cell membranes are isolated from a cell line expressing the recombinant protein, the protein is extracted from the membranes and immunoprecipitated. The  
25 proteins may then be further purified by standard protein chemistry techniques as described above.

F. Antibodies

As mentioned above, antibodies can also be used for  
30 the screening of polypeptide products encoded by the polymorphic nucleic acids of the invention. In addition, antibodies are useful in a variety of other contexts in accordance with the present invention. Such antibodies can be utilized for the diagnosis of HH and, in certain applications,  
35 targeting of affected tissues.

Thus, in accordance with another aspect of the present invention a kit is provided that is suitable for use in screening and assaying for the presence of polypeptide



products encoded by the polymorphic nucleic acids of the invention by an immunoassay through use of an antibody which specifically binds to polypeptide products encoded by the polymorphic nucleic acids of the invention in combination with a reagent for detecting the binding of the antibody to the gene product.

Once hybridoma cell lines are prepared, monoclonal antibodies can be made through conventional techniques of priming mice with pristane and interperitoneally injecting such mice with the hybrid cells to enable harvesting of the monoclonal antibodies from ascites fluid.

In connection with synthetic and semi-synthetic antibodies, such terms are intended to cover antibody fragments, isotype switched antibodies, humanized antibodies (mouse-human, human-mouse, and the like), hybrids, antibodies having plural specificities, fully synthetic antibody-like molecules, and the like.

This invention also embraces diagnostic kits for detecting DNA or RNA comprising a polymorphism of Table 1 in tissue or blood samples which comprise nucleic acid probes as described herein and instructional material. The kit may also contain additional components such as labeled compounds, as described herein, for identification of duplexed nucleic acids.

The following examples are provided to illustrate the invention but not to limit its scope. Other variants of the invention will be readily apparent to one of ordinary skill in the art and are encompassed by the appended claims.

#### EXPERIMENTAL EXAMPLES

##### I. Sequencing of 235 KB from a Homozygous Ancestral (Affected) Individual

In these studies the entire genomic sequence was determined from an HH affected individual for a region corresponding to a 235,033 bp <sup>(SEQ ID NO: 1)</sup> region surrounding the HH gene between the flanking markers D6S2238 and D6S2241. The sequence was derived from a human lymphoblastoid cell line, <sup>ATCC NO. CCL 1237</sup> HC14, that is homozygous for the ancestral HH mutation and

region. The sequence from the ancestral chromosome (Figure 2) <sup>(SEQ ID NO: 2)</sup> was compared to the sequence of the region in an unaffected individual disclosed in copending U.S.S.N. 08/724,394 (a portion of which is provided in Figure 1) <sup>(SEQ ID NO: 1)</sup> to identify polymorphic sites. A subset of the polymorphic alleles so defined were further studied to determine their frequency in a collection of random individuals.

#### A. Cosmid Library Screening

The strategy and methodology for sequencing the genomic DNA for the affected individual was essentially as described in copending U.S.S.N. 08/724,394, hereby incorporated by reference in its entirety. Basically, a cosmid library was constructed using high molecular weight DNA from HC14 cells. The library was constructed in the supercos vector (Stratagene, La Jolla, CA). Colonies were replicated onto Biotrans nylon filters (ICN) using standard techniques. Probes from genomic subclones used in the generation of the sequence of the unaffected sequence disclosed in 08/724,394 were isolated by gel electrophoresis and electroporation. Subclones were chosen at a spacing of approximately 20 KB throughout the 235 KB region. The DNA was labeled by incorporation of <sup>32</sup>P dCTP by the random primer labeling approach. Positively hybridizing clones were isolated to purity by a secondary screening step. Cosmid insert ends were sequenced to determine whether full coverage had been obtained, and which clones formed a minimal path of cosmids through the 235 KB region.

#### B. Sample Sequencing

A minimal set of cosmid clones chosen to cover the 235 KB region were prepped with the Qiagen Maxi-Prep system. Ten micrograms of DNA from each cosmid preparation were sonicated in a Heat Systems Sonicator XL and end-repaired with Klenow (USB) and T4 polymerase (USB). The sheared fragments were size selected between three to four kilobases on a 0.7% agarose gel and then ligated to BstXI linkers (Invitrogen). The ligations were gel purified on a 0.7% agarose gel and

cloned into a pSP72 derivative plasmid vector. The resulting plasmids were transformed into electrocompetent DH5a cells and plated on LB-carbenicillin plates. A sufficient number of colonies was picked to achieve 15-fold clone coverage. The appropriate number of colonies was calculated by the following equation to generate a single-fold sequence coverage: Number of colonies = size of bacterial clone (in KB)/average sequence read length (0.4 KB). These colonies were prepped in the 96-well Qiagen REAL, and the 5' to 3' DNA Prep Kit, and AGCT end-sequenced with oligo MAP1 using standard ABI Dye Terminator protocols. MAP1 was CGTTAGAACGCGGCTACAAT. (SEQ ID NO: 3)

### C. Genomic Sequencing

The MAP1 sequences from the cosmid clones HC182, HC187, HC189, HC195, HC199, HC200, HC201, HC206, HC207, and HC212 were assembled into contigs with the Staden package (available from Roger Staden, MRC). A minimal set of 3 KB clones was selected for sequencing with oligo labeled MAP2 that sits on the opposite end of the plasmid vector. The sequence of MAP2 was GCCGATTCATTAATGCAGGT. (SEQ ID NO: 4) The MAP2 sequences were entered into the Staden database in conjunction with the MAP1 sequences to generate a tiling path of 3 KB clones across the region. The plasmid 3 KB libraries were concurrently transformed in 96 well format into pox38UR (available from C. Martin, Lawrence Berkeley Laboratories). The transformants were subsequently mated with JGM (Strathman et al. P.N.A.S. 88:1247-1250 (1991) in 96 well format. All matings of the 3 KB clones within the tiling path were streaked on LB-carbenicillin-kanamycin plates and a random selection of 12 colonies per 3 KB clone was prepped in the AGCT system. The oligos -21, (SEQ ID NO: 5) CTGTAAAACGACGGCCAGTC, and REV: (SEQ ID NO: 5) GCAGGAAACAGCTATGACC (SEQ ID NO: 6) were used to sequence off both ends of the transposon. Each 3 KB clone was assembled in conjunction with the end sequence information from all cosmid clones in the region.

In some regions, the coverage of the genomic sequence by cosmids was incomplete. Any gaps in the sequence were filled by using standard PCR techniques to amplify

genomic DNA in those regions and standard ABI dye terminator chemistry to sequence the amplification products.

#### D. Identification of Polymorphic Sites

5           The assembled sequence of the cosmid clones in  
 connection with the PCR amplified genomic DNA (Figure 2) <sup>(SEQ ID NO: 2)</sup> was  
 compared to the genomic sequence of the unaffected individual  
 (Figure 1) <sup>(SEQ ID NO: 1)</sup> using the FASTA algorithm. Numeric values were  
 assigned to the sequenced regions of 1 to 235,303, wherein  
 10 base 1 refers to the first C in the CA repeat of D6S2238 and  
 base 235,303 is the last T in the GT repeat of D6S2241 of the  
 unaffected sequence (Figure 1) <sup>(SEQ ID NO: 1)</sup>. Table 1 lists the differences  
 between the two compared sequences. Note that previously  
 disclosed (Feder et al., Nature Genetics 13:399-408 (1996))  
 15 polymorphic sites D6S2238 (base 1), D6S2241 (base 235,032),  
 24d1 (base 41316), and D6S2239 (base 84841) are not included  
 in the list of new polymorphisms, although they are provided  
 for reference in a footnote to the Table and were observed in  
 the ancestral sequence. In the Table, a single base change  
 20 such as C-T refers to a C in the unaffected sequence at the  
 indicated base position that occurred as a T in the  
 corresponding position in the affected sequence. Similarly,  
 an insertion of one or more bases, such as TTT in the affected  
 sequence, is represented as "TTT INS" between the indicated  
 25 bases of the unaffected sequence. A deletion of one or more  
 bases occurring in the affected sequence, such as AAA DEL, is  
 represented as the deletion of the indicated bases in the  
 unaffected sequence.

#### 30   II. Characterization of Rare Polymorphisms

In this study about 100 of the polymorphisms of  
 Table 1 were arbitrarily chosen for further characterization.  
 Allele frequencies in the general population were estimated by  
 OLA analysis using a population of random DNAs (the "CEPH"  
 35 collection, J. Dausset et al., Genomics 6(3):575-577 (1990)).  
 These results are provided in Table 2.

One single base pair difference, occurring at base  
 35983 and designated C182.1G7T/C (an A to G change on the

opposite strand) was present in the ancestral chromosome and rare in the random DNAs. This change occurred in a noncoding region of the hemochromatosis gene near exon 7 approximately 5.3KB from the 24d1 (Cys282Tyr) mutation. OLA was used to genotype 90 hemochromatosis patients for the C182.1G7T/C base pair change. The frequency for C occurring at this position in the patients was 79.4% as compared to 5% in the random DNAs. Eighty-five of the 90 patients assayed contained identical 24d1 and C182.1G7T/C genotypes. Four of the remaining 5 patients were homozygous at 24d1 and heterozygous at C182.1G7T/C; one was heterozygous at 24d1 and homozygous at C182.1G7T/C. The primers used for this analysis were as follows.

15 PCR primers for detection:

182.1G7.F 5'-GCATCAGCGATTAAGTTCTAC -3' (SEQ ID NO:7)  
 182.1G7.R 5'-TTGCATTGTGGTGAAATCAGGG -3' (SEQ ID NO:8)

20 For the detection assay, the biotinylated primers used were as follows.

182.1G7.C 5' (b) CTGAGTAATTGTTTAAGGTGC -3' (SEQ ID NO:9)  
 182.1G7.T 5' (b) CTGAGTAATTGTTTAAGGTGT -3' (SEQ ID NO:10)

25

~~The phosphorylated digoxigenin-labeled primer used was~~ (SEQ ID NO:11)

~~182.1G7.D 5' (p) ACAAGACATAGATATGGTGG -3' (SEQ ID NO:11)~~

30 A further rare single base pair change was detected at 61,465bp. The inheritance pattern of this polymorphism, C195.1H5C/T (a G to A change on the opposite strand), is identical to that of 24d1. The frequency of T occurring at that position (C195.1H5T) observed in a set of 76 patients was 78.5% as compared to 5% in random individuals.

PCR primers for detection:

1951H5.3F 5'-GAATGTGACCGTCCCATGAG-3' (SEQ ID NO:12)  
 1951H5.3R 5'-CAACTGAATATGCAGAAAAAAGTACACC-3' (SEQ ID NO:13)

5

For the detection assay, the biotinylated primers used were:

1951H5.3.4 5' (b) AGTAGCTGGGACTCACGGTGT-3' (SEQ. ID NO:14)  
 1957H5.3.5 5' (b) AGTAGCTGGGACTCACGGTGC-3' (SEQ. ID NO:15)

10

The phosphorylated digoxigenin-labeled primer used was:

1951H5.3.6 5' (p) GCGCCACCACTCCCAGCTCAT-3' (SEQ ID NO:16)

15

These rare alleles are thus preferred surrogate markers for 24d1 and are especially useful in screening assays for the likely presence of 24d1 and/or 24d2.

All publications, patents, and patent applications cited herein are hereby incorporated by reference in their entirety.

20

BASE LOCATION	DIFFERENCE	BASE LOCATION	DIFFERENCE
47758	C-A	64788	A-G
47994	G-C	64962	G-A
48440	G-A	65891	C-T
48650	T-G	66675	G-C
48680	A-G	67186-67187	ATT INS
50240	C-T	67746-67747	TT INS
50553	G-A	68259	T-C
50586	G-T	68836	T-C
51322	G-C	68976	C-G
51747	A-G	72508	T-G
52474	C-G	72688	C-G
52733	C-A	75323-75324	T INS
52875	G-A	75887	G-C
53631-53637	TTTTTTT DEL	77519	T-C
53707	G-A	77749	G-A
54819	A-G	77908	T-C
55913	T-C	78385	C-G
56225	A-C	78592-78593	AG INS
56510	T-C	80189	T-G
56566	G-A	80279	T DEL
56618	A-T	80989-80990	A INS
57815	A-G	81193	T-C
58011	T DEL	81273	A DEL
58247-58248	T INS	82166	G-A
58926	C-G	83847	T DEL
59406	C-G	84161-84162	CA-GG
59422	G-C	84533	A-G
60221-60222	A INS	84638	T-G
60656-60657	CA DEL	85526	T-G
61162	G-A	85705	G-T
61465	G-A	86984	T-C
61607	A DEL	87655	T-C
61653	T-C	87713	A-C
61794-61795	T INS	87892	C-T
62061	G-C	88192	T DEL
62362	T-G	88528	A-G
62732	C-G	89645	A-T
63364	G-A	89728	A-G
63430-63431	GT INS	90088	T-C
63754	C-T	91193-91194	2209bp INS
63785	A-C	91373	T-C
63870-63871	A INS		

BASE LOCATION	DIFFERENCE	BASE LOCATION	DIFFERENCE
91433-91434	A INS	133572	A-C
91747	G-A	134064	T-G
93625	T DEL	136999	G-A
95116-95117	T INS	137784	C-T
96315	G-A	138903	G-A
97981	A-G	139159-139160	A INS
98351	T DEL	140359	G-A
99249	C-T	140898	C-T
100094-100095	T INS	141313	C DEL
100647-100648	TTC INS	141343	T-C
100951	C-T	142148	T-C
101610	C-G	142178	C-A
102589	C-T	142433-142434	ATAGA INS
103076-103077	TATATATATATATA INS	143783	C-T
103747	T-C	144090	C-T
105638	A-C	144220-144221	A INS
107024	C-T	144725	A-C
107322	C-T	145732-145733	AAAAAAAAAAAAAAA INS
107858	C-G	147016-147017	CG DEL
109019	A DEL	147021	G-T
109579	T DEL	147536	T-G
110021	C-A	148936	T-A
111251	C-A	149061	T-C
111425	G-A	154341	A-T
112644	T-A	154588	G-A
113001	G-C	155464	G-A
113130	C-T	158574	C-G
114026	G-A	160007	C-T
114250	A DEL	164348	A-T
115217	C-G	164499	C-G
117995	G-A	166677-166678	AAAG INS
118874	A-G	167389	G-A
119470	T-C	168506-168507	AGGATGGTCT INS
119646	G-T	168515	T-C
120853	C-T	169413-169414	AA INS
121582	G-A	170300-170301	TTGTTGTTGTTG INS
123576	A-C	170491	G-A
125581	C-T	173428	T-C
125970	G-T	173642	G-A
126197	A-G	173948	T-G
126672	A DEL	175330	T-C
126672	G-C	175836	T-C
128220-128221	A INS	176200	G-C
132569	C-T	176222	T-C



BASE LOCATION	DIFFERENCE	BASE LOCATION	DIFFERENCE
212379	G-A	231226	A-G
212637-212639	TCT DEL	231447	G-A
212696	T-C	231835	A-G
213042	T-A	232400-232402	AAA DEL
214192	A-G	232402-232403	G INS
214529-214530	TTTTTTTTTT INS	232515	T-C
214549	T-C	232703	G-T
214795	C-T	232750	A-G
214908	T-G		
214977	A-G		
215769	C-T		
215947	C-A		
216232	A-G		
217478	G-A		
219052	T-C		
219082-219083	ATATATATATATATATAT INS		
219314	C-A		
219327	G-A		
219560	C-T		
219660	C-T		
219889	G-A		
220198	G-T		
220384	G-A		
220451-220452	CAAAAA INS		
221363	G-A		
221645	G-A		
222119	T-C		
222358	A-G		
222367	A-C		
222686	A-G		
222959	T-C		
223270-223271	TT DEL		
223283	T-C		
224964	T-C		
225232	A-C		
225366-225367	TTTT INS		
225416	G-C		
225486	T-C		
226088	A-G		
228421	A-G		
230047	G-A		
230109	G-C		
230376	C-G		
230394	A-G		

\* D6S2238 occurs at base 1. 24d1 occurs at base 41316. D6S2239 occurs at base 84841. D6S2241 occurs at base 235032

## 2. Polymorphic Allele Frequencies

Location	Frequency of ancestral variant in random chromosomes	Frequency of unaffected variant in random chromosomes
232703	53%	47%
231835	53%	47%
230394	85%	15%
230376	25%	75%
230109	53%	47%
225486	45%	55%
225416	75%	25%
220198	43%	57%
219660	58%	42%
219560	53%	47%
214977	65%	35%
214908	50%	50%
214795	24%	76%
214549	53%	47%
214192	65%	35%
210299	53%	47%
208862	80%	20%
208634	48%	52%
207400	25%	75%
205284	50%	50%
204341	53%	47%
202880	58%	42%
202662	98%	2%
200027	25%	75%
199030	58%	42%
198692	55%	45%
198401	55%	45%
198055	55%	45%
195693	60%	40%
195404	25%	75%
194890	55%	45%
175330	53%	47%
173948	83%	17%
173642	55%	45%
173428	80%	20%
168515	80%	20%
160007	18%	82%
149061	58%	42%
148936	82%	18%
147536	100%	0%
147021	46%	54%
141343	55%	45%

L cati n	Fr quency f ancestral variant in random chromosomes	Frequency of unaffected variant in random chromosomes
140359	55%	45%
138903	55%	45%
132569	81%	19%
125581	18%	82%
121582	80%	20%
120853	18%	82%
118874	85%	15%
115217	50%	50%
113130	40%	60%
113001	48%	52%
107858	48%	52%
103747	50%	50%
96315	25%	75%
91194	80%	20%
90088	75%	25%
89728	50%	50%
89645	50%	50%
88528	63%	37%
87892	75%	25%
87713	60%	40%
87655	50%	50%
86984	79%	21%
85705	50%	50%
85526	50%	50%
84638	50%	50%
84533	50%	50%
82166	78%	22%
81193	58%	42%
80189	50%	50%
78385	80%	20%
77908	88%	12%
68976	50%	50%
68259	51%	49%
66675	80%	20%
62732	50%	50%
62362	40%	60%
61653	48%	52%
61465	5%	95%
61162	60%	40%
53707	100%	0%
52875	50%	50%
52733	74%	26%

Location	Frequency of ancestral variant in random chromosomes	Frequency of unaffected variant in random chromosomes
52474	47%	53%
50586	50%	50%
50553	50%	50%
50240	50%	50%
48680	53%	47%
48650	63%	37%
48440	50%	50%
47255	50%	50%
46601	53%	47%
45567	49%	51%
41316	5%	95%
40431	20%	80%
38526	23%	77%
37411	70%	30%
35983	5%	95%